Communauté UNIVERSITÉ Grenoble Alpes

Rapport d'évaluation du mémoire de thèse / Evaluation report of the PhD thesis

Doctorant	Nom prénom / Full name	[Javier Olivares]
PhD student	Ecole Doctorale / Doctoral School	[Astrophysics]

Titre thèse / PhD Title [Bayesian Hierarchical Modelling of Young Stellar Clusters]

Rapporteur	Nom prénom / Full name	[Coryn Bailer-Jones]
Reviewer	Etablissement / Institution	[Max Planck Institute for Astronomy, Heidelberg]
	Statut, fonction / Status, position	[Senior staff member]

Qualité du mémoire, rédaction & illustrations / Thesis quality, style & illustrations

⇨ Satisfaisant / Satisfactory [ ] Bon / Good [ ] Très bon / Very good [X] Exceptionnel [ ]

Commentaires/comments :

[
see main report below
]

Contexte, état de l'art, collaborations / Background, state of the art, collaborations :

Commentaires/comments :

[
Some of this work has been done in collaboration, at some level, with the DANCe project. In particular, the input data were obtained through collaborative effort and/or are already published. Some of the work in this thesis has been submitted as a multi-author publication to an international journal: Olivares et al. 2017, A&A submitted. The thesis itself otherwise does not indicate that it includes work done by others.
]

Qualité scientifique, méthodologie, expérimentations, validation
Scientific quality, methodology, experiments, validation

⇨ Satisfaisant / Satisfactory [ ] Bon / Good [ ] Très bon / Very good [X] Exceptionnel [ ]

Commentaires/comments :

[
see main report below
]

Apports personnels, originalité, valorisation, perspectives
Personal contributions, originality, valorization, prospects

Commentaires/comments :

[
see above comments on "collaboration"
]

Conclusions du rapporteur / Reviewer's conclusions
Commentaires/comments :

Overview

This thesis is concerned with using photometric and astrometric data to infer the properties of open clusters, in particular the luminosity and mass functions. Chapter 2 provides a nice review of previous work on the Pleiades, to which the method is applied. Chapter 3 describes the development of a partially hierarchical probabilistic model for combining the different types of data and for accommodating unknown membership status of the observed targets. Section 4 describes the results of applying the method to the Pleiades, and compares the results with previous publications.

Building on earlier work, this thesis appears to be a very thorough investigation, which has carefully considered several complicating factors neglected in the past. A few treatments are inevitably less detailed (e.g. binarity, completeness), but this thesis makes an important contribution to efforts to understand clusters. It is a valuable and illustrative application of Bayesian hierarchical models in astronomy. The method and results are mostly very clearly presented, especially in the earlier parts of the thesis, although in places the exact procedure was less clear and here a bit more top-level explanation would have helped guide the reader through the details.

In my comments below, page numbers refer to v1 of the thesis. (Many comments are asked as questions for further consideration, but neither a written reply nor specific changes are expected.)

Avis du rapporteur / Reviewer's opinion :
Défavorable à la soutenance / Unfavorable to the defence [ ] Favorable [X]

Date [24 August 2017] Signature [C. Bailer-Jones]

Visa du directeur de l'école doctorale :

Rapport détaillé, commentaires libres, questionnements, correction demandées
Detailed report, free comments, questions, requested corrections

General comments

It's good to see that you take (unknown) binarity into account in your cluster model, but why only equal-mass binaries? Exactly zero systems will have equal masses, so why not relax this by adding an additional parameter (the mass ratio)? I don't suppose it would be a big additional computational burden.

I did not understand the procedure for dealing with missing data. Does it vary? Section 3 suggests a marginalization over them (but see my specific comments below), whereas later comments indicate that different things were done at different times (e.g. nearest neighbours interpolation on p. 137).

One of the motivations for the hierarchical approach is to reduce the dependence on the choice of priors, by introducing one or more levels of marginalization. Ultimately, of course, one must set top-level priors (whether empirical or theoretical, parametric or nonparametric, this makes no difference). In reading through the specification of the many priors in chapter 3, I found myself asking how sensitive the results are to these choices (and there are a lot of choices in Table 3.2). It would be a lot of work to do a sensitivity test, but it would have been interesting to see how different the results would be from some plausible non-hierarchical model. How much do the point estimates (and in particular the confidence intervals) on the parameters of interest change by taking a (simpler) non-hierarchical approach?

I am missing a clear exposition of the big picture of the photometric model for the cluster. I understand that each photometric band is modelled as a spline function of the colour index CI, with intrinsic multivariate Gaussian scatter. (The GMMs are not used for the photometric model?) Do you derive a posterior PDF over the magnitudes for each star (and these are the inputs for determining the LF)? Or do you infer posteriors only for the parameters of the model?

Given N photometric bands, you choose to work with one colour index and N-1 magnitudes?
Is there any particular reason for this rather than, say, N-1 colour indices and 1 magnitude?

Bold and non-bold typefaces for mathematical symbols are sometimes used interchangeably, which is confusing, e.g. for theta, phi, and pi on p. 60. (And on page 61 subscripts suddenly appears on theta and phi, even though they are unchanged.)

Specific comments

p. 59: If Fig. 3.2 is showing the distributions where at least one other photometric band is missing, it would have been nice to see the distributions for objects with no missing data (and then not to scale the distributions to have a maximum of 1.) Is it really the case that more than 99% of the data used in the model have at least one photometric band missing (one interpretation of the statement in the middle of p. 112)?

p. 59, eqn. 3.20: This seems to be incorrect. The integral is of P(d|theta) over just one component of d, namely x_i. In that case the result cannot be P(d|theta), but P(d'|theta), where d' is d with the element i removed (marginalized over). The result of a marginalization cannot be a PDF in what was marginalized over. There seems to be some missing explanation here, and the approach to missing data in this work remains unclear to me. (Just after eqn. 3.26 it says that missing values remain after all. But in section 4 there is a comparison with synthetic data sets which do and do not have missing data, so something presumably was done.)

p. 60, eqn. 3.23: What is the justification/need for replacing the N-dimensional integral with an N-dimensional sum? Is it because each q_n can only be 0 or 1 (in which case what is delta_q?). A couple more lines explanation would have been useful here.

p. 61: The first sentence of the final paragraph of section 3.3.2 doesn't make sense (and doesn't help me understand how you have treated missing data). In fact I don't really understand the rest of the paragraph. (Also, does "belong to the multivariate normal family" just mean "are multivariate normal"?)

p. 62: Including the few expected Pleiades members in the fit to the field model is pragmatic, and certainly plausible for the 3D spatial distribution because the Pleiades is not very dense. But the cluster stands out more in the HRD and kinematics, so might it not create some bias here? After all, it's not just the number of "contaminating" objects which counts, but how different their distribution is. I can nonetheless imagine that the impact is small. Was this tested?

sec 3.3.3: Some plots of the GMM fits to the photometry of the field would have been interesting to see.

sec. 3.3.4 (see also "general comments"):
- Early on p.66 it's not clear to me in what space the splines are being fit. The parameter vector "t" is later noted to be a vector of the data, but then I don't see how only one knot per dimension (quoted on p. 66) can be specified.
- What about non-equal-mass binaries?
- I have to confess that I find this section a bit confusing and lacking the clarity of earlier sections.

p. 72: Do you really use RA and Dec (as suggested here) rather than tangent plane coordinates (as stated on p. 76)?

p. 85: At the risk of being pedantic, I would say that (i) PSO is a maximizer of functions, not a MAP finder, and (ii) no optimizer is guaranteed to find the global optimum for a nonlinear model. The issue we face is not the inability to find the global maximum - all we can ever go is converge on something we accept as a good optimum - but how flexible the optimizer is in escaping local optima.
In a similar vein, it is not correct that the MCMC can find the true distribution. With a finite number of function evaluations you only ever get an approximation.

sec. 3.7.2. I think you mean Daniel Foreman-Mackey rather than David Foreman.

p. 108: Are you sure that a "random" classifier has an AUC of 0.5? Or rather, what do you really mean by a "random classifier"? I would define this as one which assign classes with your prior probabilities (which I understand from Table 3.2 to be p_field=0.98 and p_cluster=0.02). Even if I assigned all stars as field stars, 98% of these assignments would be correct, which also sounds good. I'm not doubting the quality of the classifier. My point is rather that a bit more explanation would have been nice to put an AUC of 0.992 into a more meaningful context.

sec. 4.2: I am assuming that all results from here on concern the real data, as opposed to the synthetic data. I also think it would also have been more natural to present the results on the real data first, and then to compare them with other results.

sec. 4.2.1: I presume these cross matches took into account epoch differences (done by the CDS tool) and (where available) proper motions?

Fig. 4.7: A casual glance shows that BHM is more optimistic about cluster membership than Bouy is (I am referring to the thick stripe at low values of Bouy's probability for the full range of BHM probability). Might this be a result of how probabilities are (implicitly) calibrated, or could it be something more fundamental in the data/method, e.g. potential biases?

p. 112: I don't understand the origin of the statements in the middle of this page concerning the consequences of missing data. It is stated that when data are missing, the Bouy likelihood ratios (cluster to field) are larger. Why? When data fields are missing for a star, the likelihood has to be rescaled to accommodate this, otherwise you cannot combine/compare the likelihoods for different stars (they would have different units, for example). Why would this lead to one model (cluster or field) being favoured?

sec. 4.5: Am I right in thinking that this section has nothing to do with the BHM? I ask, because it is stated in the first paragraph that uniform priors are used for all parameters. Why now switch to uniform priors? The Bayesian evidence is generally more sensitive to the choice of priors than is the posterior, so this strikes me as a very odd thing to do, especially given the careful discussion of priors for the BHM. How have the bounds of the priors been chosen, and how sensitive are the results to these choices? Also, why not retain the hierarchical approach here? (Too computationally demanding?)

sec. 4.5.2 (last paragraph): If there's no evidence for ellipticity, isn't the model then radially symmetric?

sec. 4.7 and 4.8: Do the derivations of the LF and MF use just a single point estimate of the magnitudes for each star? Or just the inferred parameters of the photometric model? (See also my comment below on Fig. 4.22)

sec. 4.7.1: "To derive the J, H, Ks magnitude distributions, I first transform the true CI distribution into the J, H, Ks apparent magnitude distributions." When reading this I ask myself "as CI=i-Ks, how can a distribution on this be transformed into a distribution of J or H?" I see this later from the maths: you have additional probabilistic dependencies, so this isn't really the transformation you mention; P (CI|...) is just one ingredient. A brief conceptual overview, before diving into the maths, might have helped the reader here.

p. 134/135: "to obtain the absolute magnitude distributions, I subtract this parallax distribution, by means of a convolution, to the J, H, Ks magnitude distribution". You presumably mean you "subtract" the distribution of the *log* of the parallaxes from the magnitude distributions. The approach taken here is rather indirect: the "fully" Bayesian approach would instead infer the luminosity distribution from the data directly. I presume you eschewed this approach for computational reasons?

p. 135: I don't follow why the maximum density in the i,Ks diagram should (must?) be selected as the "upper completeness limit". And what does that term mean? Does it mean that the sample is 100% complete for brighter sources (up to the "lower completeness limit")? I don't see why that would be (unless you are making some very simple, but unstated, assumptions). Also, what is the justification for the value of the lower completeness limit?

Fig 4.22: I missed how the "BHM" results differs from the "samples" results. Both are from this work, but the former only uses the candidates found by the BHM. So which objects were used for the latter?

Fig. 4.22: The differences between the Bouy et al. and your determinations of the LF appear quite small (although the vertical axis is a log scale, so maybe their integrals over the "complete" regions show larger differences; it would have been interesting to report this). Should we therefore conclude that the simpler treatment in Bouy et al. was adequate for the LF determination?

Table 4.8: I don't follow the explanation of the evidence computation. The text suggests the parameters are fixed (at the MAP), and then the likelihood is computed. This is not the evidence, however.

sec. 4.8: I wouldn't say Trapezium and Hyades are just the Pleiades at different times. I don't think all clusters are born identical!

]